Centrality measures and thermodynamic formalism for complex networks

Delvenne, J.-C.; Libert, A.-S.

Published in:
Physical Review E - Statistical, Nonlinear, and Soft Matter Physics

DOI:
10.1103/PhysRevE.83.046117

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Centrality measures and thermodynamic formalism for complex networks

Jean-Charles Delvenne[*,†] and Anne-Sophie Libert[‡]
Department of Mathematics, Namur Centre for Complex Systems, Facultés Universitaires Notre-Dame de la Paix, B-5000 Namur, Belgium

(Received 27 September 2010; revised manuscript received 4 February 2011; published 22 April 2011)

In the study of small and large networks it is customary to perform a simple random walk where the random walker jumps from one node to one of its neighbors with uniform probability. The properties of this random walk are intimately related to the combinatorial properties of the network. In this paper we propose to use the Ruelle-Bowens random walk instead, whose probability transitions are chosen in order to maximize the entropy rate of the walk on an unweighted graph. If the graph is weighted, then a free energy is optimized instead of the entropy rate. Specifically, we introduce a centrality measure for large networks, which is the stationary distribution attained by the Ruelle-Bowens random walk; we name it entropy rank. We introduce a more general version, which is able to deal with disconnected networks, under the name of free-energy rank. We compare the properties of those centrality measures with the classic PageRank and hyperlink-induced topic search (HITS) on both toy and real-life examples, in particular their robustness to small modifications of the network. We show that our centrality measures are more discriminating than PageRank, since they are able to distinguish clearly pages that PageRank regards as almost equally interesting, and are more sensitive to the medium-scale details of the graph.

DOI: 10.1103/PhysRevE.83.046117 PACS number(s): 89.75.Hc, 05.40.Fb, 89.70.Cf, 89.75.Fb

I. INTRODUCTION

In the past decade a tremendous amount of data on how various agents interact with each other has been collected. This can be people exchanging phone calls in sociology, web pages pointing to each other through hyperlinks, genes influencing the expression of other genes in genetics, food webs in ecology, etc. These large to huge graphs require powerful methods of analysis in order to identify the key structures of the graph. A particular problem retains our attention here: centrality measures.

One of the most prominent application of centrality measures is the web search, where the most central, best connected pages through the network of hyperlinks are often the most relevant regarding their content. Google and other web search engines attribute to each page of the web a PageRank score, which measures how well connected the page is with respect to other pages [1]. More specifically, a page has a high PageRank if pointed to by pages with a high PageRank. Kleinberg [2] proposed the hyperlink-induced topic search (HITS) method, where a page is a good hub on a topic if it points to good authorities on this topic and a page is a good authority if pointed at by good hubs. Other variants have been proposed by several authors; we mention only Ding *et al*. [3], who proposed a framework generalizing HITS and PageRank, and Akian *et al*. [4], who used thermodynamic concepts in a different way from us. Those methods are all variants of the earlier eigenvector centrality method [5], which computes the dominant left eigenvector of the adjacency matrix as the centrality measure. Other centrality measures, such as betweenness and closeness, based on counting the shortest paths between nodes, have been popular [6]. Although the web now constitutes the most spectacular application of centrality measures, they were first used for social network analysis [5] and have found many other applications, most recently in economic networks [7,8].

In this paper we apply methods from Ruelle's thermodynamic formalism to the field of large graphs; in particular we introduce entropy rank and free-energy rank methods, which rank the nodes of a network. Let us consider a strongly connected graph. While PageRank is based on the simple random walk, where a random walker jumps from one node to any of its d out-neighbors with uniform probability $1/d$, entropy rank is based on the Ruelle-Bowens random walk [9,10]. This random walk on the graph obeys transition probabilities that are chosen to make all paths of the same length occur with approximately equal probability. In other words, the transition probabilities of the Ruelle-Bowens random walk are chosen so as to maximize the entropy rate of the random walk. Entropy rank is now defined as the stationary distribution of the Ruelle-Bowens random walk.

If the graph is not strongly connected, the entropy rank will have undesired effects, or even will not be uniquely defined. In this case, we use a trick close to PageRank's teleportation trick. Given any network, one may complete the graph with all the nonedges and assign them a certain constant weight. We now have a weighted complete graph, with two different values for the edges. The Ruelle-Bowens random walk is also defined for weighted graphs, where the weights are interpreted as energies. Instead of maximizing the entropy rate of the random walk, we maximize the sum of the entropy rate with the average energy of the edge; this sum is called the free-energy rate of the random walk. As a result, the random walk will have a tendency to visit high-energy edges more often (it should be noted that Ruelle's sign convention for energy, which we follow here, is opposite that of most physicists, who usually consider low energy to be more probable). The stationary distribution of the

[*]jean-charles.delvenne@uclouvain.be
[†]Present address: Department of Applied Mathematics, Université Catholique de Louvain, B-1348 Louvain-la Neuve, Belgium.
[‡]anne-sophie.libert@fundp.ac.be

Ruelle-Bowens random walk on the complete weighted graph is what we call the free-energy rank.

In the undirected case, the entropy rank and free-energy rank essentially coincide with eigenvector centrality. In the directed case, they share attributes with PageRank and HITS. For example, entropy and free energy ranks attribute high scores to nodes that point to high-score nodes or are pointed at by high-score nodes, while only the latter is of direct relevance for PageRank.

We look at a toy example and a 289 000-node piece of the web to examine the ability of the free-energy rank to better discriminate between the nodes. For the toy example we notice that nodes identically ranked by PageRank are distinguished as different by the entropy and free-energy ranks. In the large size example, we notice that the distribution of centrality scores is more uneven for the free-energy rank than for PageRank. It is thus better at separating central from noncentral nodes. Moreover, we introduce groups, all nodes of which point to a single page, in order to see how the ranking of this page is enhanced. We observe that the free-energy rank is more sensitive to such perturbations than PageRank.

The goal of this paper is therefore to introduce alternative centrality measures and, more generally, to illustrate and promote the use of the Ruelle-Bowens random walk for complex networks. Although Ruelle's thermodynamic formalism, based on various powerful generalizations of Ruelle-Bowens random walks, is a physics-inspired, mathematically profound theory, it has received little attention so far in the study of large graphs and complex networks. The connection between the Ruelle-Bowens random walk and the robustness of a complex network is explored in Ref. [11]. Since an earlier version of the present article appeared [12], several papers have computed the entropy rate for the simple random walk, the Ruelle-Bowens random walk, and others on a variety of synthetic and real-life networks [13,14]. In Ref. [15] Sinatra *et al.* show how to implement approximately the Ruelle-Bowens random walk using only local information.

Many algorithms proceed by performing a simple random walk on the graph in order to extract some combinatorial features. It has been shown in the area of community detection that different variants of the simple random walk (e.g., discrete-time or continuous-time random walks) are able to highlight different features of a complex network [16,17]; in addition, the entropy rate of a simple random walk is used as a fundamental tool in Ref. [18] to uncover communities. How the Ruelle-Bowens random walk can help in the understanding of complex networks is an almost pristine field of research.

II. PAGERANK

A. PageRank: First approach

We now discuss the principle of PageRank. PageRank can be defined in any kind of network, as mentioned in the Introduction. Nevertheless, we will take as an example the case of the web graph, with pages as nodes and hyperlinks as edges. Imagine a surfer starting from a page and clicking randomly on the hyperlinks on the page, each with equal probability. By repeating this process indefinitely, one may compute the asymptotic stationary probability distribution of the surfer. By elementary Markov chain theory, this distribution exists and does not depend on the initial state if the graph is strongly connected and aperiodic. It is given by the dominant left eigenvector of the row-stochastic, normalized adjacency matrix of the graph $D^{-1}A$. Here the adjacency matrix A is defined by $A_{ij} = 1$ if there is an edge from i to j and $A_{ij} = 0$ otherwise, and D is the diagonal matrix of outdegrees. In the strongly connected aperiodic case the distribution is also the vector of frequencies at which every node is visited by the random surfer. The PageRank [1] is then defined as this stationary distribution.

The problem with this definition is that many graphs of interest, including the web graph, are not strongly connected. In particular, many pages contain no hyperlink or are the target of no hyperlink. An improvement is therefore needed.

B. PageRank with teleportation

To overcome this problem, the possibility is given to the random surfer, with some probability $0 < 1 - \alpha < 1$, to jump to any other page of the web (with uniform distribution). The surfer follows a hyperlink of the current page with probability α. If there is no hyperlink, then the surfer jumps to a random page with probability 1 (which we may call a teleportation, as this jump is not local).

Let $\tilde{A}$ be the adjacency matrix of the graph, with every nonzero row normalized to 1. Then the stochastic matrix P describing the Markov chain is constructed as follows. Let e be the vector of all 1's, normalized in order to sum to 1. The ith row is equal to $(1-\alpha)e^T + \alpha\tilde{A}_i$ if $\tilde{A}_i$ (the ith row of $\tilde{A}$) is nonzero. If $\tilde{A}_i = 0$ then the ith row is taken as e^T. The left dominant eigenvector of this matrix P, normalized in order to sum to 1, gives the unique stationary distribution on the vertices. The PageRank is now defined as this stationary distribution. Note that in practice, the entries of e are not necessarily all equal but can be chosen in order to favor some pages.

If α tends toward 1, then we recover the first approach above (provided the graph is aperiodic and strongly connected). If α tends toward 0, then the stationary distribution tends toward the uniform distribution. For all $\alpha < 1$, the PageRank is well defined on all graphs.

PageRank has demonstrated its power in applications on the web and elsewhere. However, we might argue that it may fail to distinguish the most interesting nodes in some cases.

TABLE I. PageRank, free-energy rank, and entropy rank for the network of Fig. 1.

Vertex	PageRank ($\alpha = 1$)	PageRank ($\alpha = 0.9$)	Entropy rank	Free-energy rank ($E = 0.03$)
1	0.1705	0.1549	0.2464	0.2400
2	0.2045	0.1965	0.2487	0.2458
3	0.1818	0.1644	0.2487	0.2460
4	0.1705	0.1549	0.2464	0.2400
5	0.0909	0.1035	0.0032	0.0099
6	0.0455	0.0601	0.0001	0.0019
7	0.0909	0.1057	0.0032	0.0076
8	0.0455	0.0601	0.0031	0.0087

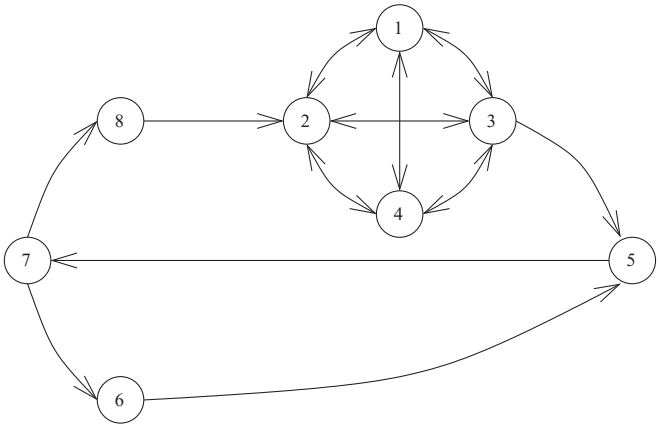

FIG. 1. Toy network. Ranking scores computed according to different methods (see Table I). Vertices 6 and 8 have the same PageRank for whatever value of α is chosen, while both entropy rank and free-energy rank are able to distinguish them. The gap of entropy rank between the best vertices and worst vertices is larger than for any other method.

Indeed, let us take the graph of Fig. 1 (see also Table I). Vertices 1, 2, 3, and 4 form a complete directed subgraph, hence they concentrate most of the probability for values of α close to 1, as expected; however, they attribute an equal probability to 6 and 8, as we may be easily convinced. This might be argued as intuitively undesirable because 8 is obviously a better page than 6: It directly points to the most interesting pages.

III. ENTROPY RANK

We now introduce a centrality measure that we call the entropy rank. We assume again a surfer on a strongly connected, aperiodic graph, but instead of choosing a hyperlink with equal probabilities $1/d$, it chooses the first hyperlink on the page with a specific probability $p > 0$, the second with a probability $p' > 0$, etc. We want to choose those probabilities in order to make the long-term behavior of the surfer as unpredictable as possible; in other words, we want all the possible paths of the surfer to be (almost) equally probable.

Let us be more specific. Assume that on every page i we have chosen the probability $p_{ij} > 0$ of a transition toward page j if there is a hyperlink from i to j and the probability $p_{ij} = 0$ if there is none. This will result in the surfer being asymptotically in every state i with a certain stationary probability π_i. The vector π is the dominant left eigenvector of the row-stochastic matrix $P = (p_{ij})$. We may then compute the probability of the random surfer following the path $ijk \ldots mn$ as $\pi_i p_{ij} p_{jk} \ldots p_{mn}$. For every t we may define the Shannon entropy $H(t)$ of all paths of length t that the random surfer can follow. Then the entropy rate of the random surfer is defined as $\limsup_{t\to\infty}[H(t)/t]$ (see, e.g., Ref. [19]). This entropy rate depends of course on the transition probabilities p_{ij}. We now choose the entries p_{ij} in order to maximize the entropy rate of the random surfer. We now show how to compute p_{ij} and the resulting entropy from the adjacency matrix A of the graph.

The Shannon entropy of a probability distribution over a set of N elements is at most $\log N$ and the uniform distribution is the only distribution to achieve this bound, as is well known. Now consider a probability distribution of a random variable X that is uniform up to a factor a, meaning that the probability of any event is at most a/N. Then the Shannon entropy of this distribution is the convex combination of terms $-\log \mathrm{Prob}(X = i)$, each of which is at least $\log N - \log a$. Hence the Shannon entropy itself is at least $\log N - \log a$.

For any probability distribution over the paths, the Shannon entropy of paths of length t is at most $\log |\{\text{paths of length } t\}|$. Hence the entropy rate is at most $\limsup_{t\to\infty} [\log |\{\text{paths of length } t\}|/t]$. This quantity is called the topological entropy of the graph because it is not dependent on any particular probability distribution but is intrinsic to the graph. Since the number of paths of length t is the sum of all entries of A^t, the topological entropy is readily seen to be equal to the logarithm of the spectral radius of the adjacency matrix A.

Now, following Parry [9], we exhibit a particular probability distribution whose entropy rate is precisely the topological entropy of the graph. Let λ be the dominant eigenvalue of A of maximal magnitude, u be a non-negative left eigenvector for λ, and v be a non-negative right eigenvector for λ. We thus have $u^T A = \lambda u^T$ and $Av = \lambda v$. Their existence is ensured by the Perron-Frobenius theorem and they can be computed by the power method. We normalize u such that $\sum_i u_i = 1$ and v such that $\sum_i u_i v_i = 1$. We choose the probability p_{ij} to take the existing edge (i, j) starting from i to be

$$p_{ij} = v_j/\lambda v_i. \tag{1}$$

This is indeed a probability distribution over the outgoing edges of i since $\sum_{j:(i,j)\text{is an edge}} v_j/\lambda v_i = \sum_j A_{ij} v_j/\lambda v_i = \lambda^{-1}(Av)_i/v_i = 1$. Then the row-stochastic transition matrix is

$$P = \lambda^{-1}\mathrm{diag}(v)^{-1} A\ \mathrm{diag}(v), \tag{2}$$

where $\mathrm{diag}(v)$ is the diagonal matrix formed from vector v.

The distribution attributing a probability

$$\pi_i = u_i v_i \tag{3}$$

to vertex i is an invariant distribution on the vertices of the Markov chain. Indeed, $\pi^T P = \pi^T \lambda^{-1} \mathrm{diag}(v)^{-1} A\ \mathrm{diag}(v) = u^T \lambda^{-1} A\ \mathrm{diag}(v) = u^T \mathrm{diag}(v) = \pi^T$.

The probability of path ij is $u_i v_i \lambda^{-1} v_j/v_i = \lambda^{-1} u_i v_j$, the probability of path ijk is $\lambda^{-1} u_i v_j \lambda^{-1} v_k/v_j = \lambda^{-2} u_i v_k$, and, more generally, any path of length t going from vertex i to vertex j has a probability $\lambda^{-t} u_i v_j$ (which does not depend on the intermediate vertices). We know that the number of paths of length t is on the order of λ^t (up to a factor). Hence the probability distribution over paths of fixed length is uniform up to a factor (which does not depend on t). The Shannon entropy of paths of length t therefore grows as $t \log \lambda$, up to an additive constant. The entropy rate of this distribution is thus $\log \lambda$, which is optimal.

In brief, we have proved the following: The behavior of a random surfer with maximal entropy rate can be computed from a left and a right non-negative dominant eigenvector, obtained, for instance, with the power method, and the resulting distribution on vertices is given by the component-related product of the two eigenvectors.

Definition 1. The entropy rank of vertex i of an unweighted, strongly connected, aperiodic graph is defined as the probability $u_i v_i$, where u (v) is the left (right) dominant eigenvector of the adjacency matrix.

Since the graph is strongly connected and aperiodic, then λ, u, and v are unique and positive according to the Perron-Frobenius theorem. The entropy rank is then uniquely defined and nonzero on every vertex. Note that the matrix $\lambda^{-t} A^t$ can be shown to converge to vu^T, whose diagonal gives the vertex probability distribution. As shown in Ref. [9], when the graph is strongly connected there is no other probability distribution that maximizes the entropy rate. (See the numerical example of Fig. 1.) A more trivial example is the complete graph on n vertices, for which A is the matrix of 1's (except on the diagonal); we see that the entropy rate has the maximal value $\log(n-1)$ for the uniform distribution.

Note also that if we reverse all edges of the graph, then the matrix A is replaced by A^T, the vectors u and v switch their roles, and the final value for the entropy rank is the same. Hence the entropy method takes into account not only the paths leading to a vertex, but also the paths issued from a vertex. In the case of an undirected graph, as both eigenvectors are identical, the entropy rank provides the same ranking as eigenvector centrality, which ranks nodes according to their entry of the left eigenvector.

IV. FREE-ENERGY RANK

We want a method that gives every graph, even those not strongly connected, a unique centrality score of the nodes that is nonzero on every vertex. That is why we add the following improvement, which is a particular case of Ruelle's thermodynamic formalism [10]. On the complete directed graph with self-loops that extends the original graph we attribute an energy $U = 0$ to the edges of the original graph and an energy $U = -U_0 < 0$ to the other edges. Now consider the set of all paths in the complete graph. The energy of a path is defined as the energy of its first edge; the reason for that is because the first edge of the path represents the current transition, which determines the current energy. On this set we want to put an invariant probability measure that maximizes the quantity $S + \overline{U}$, where S is the entropy rate and $\overline{U}$ is the expected energy for the probability measure; in other words, it is the ensemble average of the energy. The maximum of this quantity is analogous to what is called free energy in thermodynamics (up to a simple change of variable, since it usually appears in thermodynamics under the form $\overline{U} - TS$, for some temperature T); more precisely, we will call it a free-energy rate because S is an entropy rate rather than a Shannon entropy. It is also called topological pressure in the literature of thermodynamic formalism.

This time we consider the matrix B such that $B_{ij} = \exp(U_{ij})$, where U_{ij} is the energy of the edge ij. Note that if $U_0 \to \infty$, then B converges to the adjacency matrix A. Note also that the matrix B can be obtained from A by replacing zero entries with e^{-U_0}.

It is possible to see that the optimal set of transition probabilities exists and is unique; we can compute it in the following way. Let λ, u, and v be such that λ is the dominant eigenvalue of B, $u^T B = \lambda u^T$ (left eigenvector), $Bv = \lambda v$ (right eigenvector), $\sum_i u_i = 1$ $u > 0$, and $\sum_i u_i v_i = 1$ $v > 0$. These objects exist and are unique according to the Perron-Frobenius theorem.

Now we claim that the set of transition probabilities optimal for the free-energy rate attributes a probability of

$$\pi_i = u_i v_i \tag{4}$$

to be in vertex i. It also attributes a probability

$$p_{ij} = \lambda^{-1} \exp(U_{ij}) v_j / v_i \tag{5}$$

to the transition $i \to j$ of energy U_{ij}. In matrix notation, the row-stochastic matrix of transition probabilities is written

$$P = \lambda^{-1} \mathrm{diag}(v)^{-1} B \, \mathrm{diag}(v). \tag{6}$$

These claims can be derived as corollaries to Ruelle's more general results [10], but we prefer to give an elementary argument for the sake of self-containedness.

Definition 2. For a given $U_0 > 0$, the free-energy rank of vertex i of an unweighted directed graph is defined as the probability

$$u_i v_i, \tag{7}$$

where u (v) is the left (right) dominant eigenvector of the matrix B obtained from the adjacency matrix by replacing the 0 entries with e^{-U_0}.

The proof of the claim, which we give for the sake of clarity, relies on the following result, which is well known in statistical physics (see, for instance, Ref. [10]). Given a finite set endowed with a real-valued energy function, the only probability distribution on this set that maximizes the free energy (here the sum of the Shannon entropy and the expected energy) is the Boltzmann distribution, which attributes probability $\exp(U_i)/\sum_i \exp(U_i)$ to element i. The free energy is then $\log \sum_i \exp(U_i)$. If a probability distribution is the Boltzmann distribution up to a factor a, meaning that the probability for element i is at most $a \exp(U_i)/\sum_i \exp(U_i)$, then the corresponding free energy is at least $\log \sum_i \exp(U_i) - \log a$.

The random walk described just above gives a probability $\lambda^{-t} \exp(\sum U_{kl}) u_i v_j$ to a path of length t from vertex i to vertex j, where $\sum U_{kl}$ is the sum of energies of all edges kl on the path. This has the form of a Boltzmann distribution, up to a factor. Now if we give to a path of length t a total path energy that is the sum of all energies of its t individual edges, then this probability distribution yields a total path free energy equal to $\log \sum_{\text{paths of length } t} \exp(\sum U_{kl})$ up to an additive constant (independent of t), which is almost maximal. This total path free energy, divided by t, gives for $t \to \infty$ a maximal free-energy rate $S + \overline{U}$. Note that the expected total path energy of a path of length t is exactly $t\overline{U}$. Note also that the maximal free-energy rate is again $\log \lambda$, the logarithm of the spectral radius of B.

The interpretation of this framework is as follows. A random surfer can jump from any page to any page, with an energy cost of U_0 if no hyperlink is present between the pages. The surfer, whose aim is to optimize the free-energy rate $S + \overline{U}$, is therefore incited to follow hyperlinks (edges of the graph) in priority. If the energy gap U_0 is 0, then the optimal probability is uniform. If the energy gap is high, then the surfer is incited to follow hyperlinks most of the time. Such

a phenomenon is similar to what is observed when varying the factor α between 0 and 1 in the PageRank method (as detailed in Sec. II B).

One may ask how to choose a reasonable U_0. With the knowledge that $\alpha = 0.85$ or 0.9, for instance, works well in the case of PageRank, one may develop a heuristic argument to find a corresponding value of U_0 as follows. Suppose that the outdegree of the graph is constant. Then the right eigenvector of A is constant as well and the PageRank (for $\alpha = 1$) is equal to the entropy rank if it is well defined. Moreover, for every value of α there is a corresponding value of U_0 such that the PageRank and the free-energy rank coincide. A calculation shows that this value is such that

$$E = e^{-U_0} = 1 \Big/ \left(1 + \frac{\alpha N}{(1-\alpha)d}\right). \tag{8}$$

Indeed, for the adjacency matrix A on N nodes, the simple random walk and the Ruelle-Bowens random walk coincide, with the transition matrix $P = \frac{1}{d}A$. The PageRank for any α is the left eigenvector of $\alpha P + (1-\alpha)\frac{1}{N}\mathbf{1}$, where $\mathbf{1}$ is the $N \times N$ matrix of ones, and the free-energy rank is the left eigenvector of $(1-E)A + E\mathbf{1}$. The imposition of the equality between PageRank and the free-energy rank leads to the formula above [Eq. (8)].

When the graph is not with constant outdegree, PageRank and entropy rank do not coincide in general. However, we may replace d as the average outdegree $\langle d \rangle$ to guess a reasonable value for E corresponding to a given value of α.

The free-energy method also gives a nonzero probability to any vertex of the graph. An example of such a calculation is shown in Fig. 1. We consider a value of U_0 that is equivalent to $\alpha = 0.9$. We note that, this time, node 7 has a free-energy rank lower than node 8, which indicates that page 8 is more interesting, which is a sensible claim. Again, the free-energy rank is invariant under a reversal of edges.

V. NUMERICAL EXPERIMENTS

We now compare the PageRank and free-energy rank distributions for a 289 000-node piece of the Stanford web [20]. The distributions are shown in Fig. 2. The top panel indicates the PageRank scores associated with $\alpha = 0.9$ and shows the well-known power-law trend of the PageRank distribution. The free-energy rank distribution with an equivalent value of U_0 is represented in the middle panel. While the qualitative behavior is similar, the values of the free-energy rank are more spread out: the ratio of the centrality score between the best and worst pages is much higher when the centrality score is a free-energy rank rather than a PageRank. Finally, for smaller value of U_0 (or, equivalently, larger value of E), the main pages are highlighted in the distribution (bottom panel of Fig. 2), while the worst pages are gathered to the same free-energy rank value. This is not surprising since the distribution is determined by the left and right eigenvectors of a matrix B whose entries are all 1 or E. Therefore, the right eigenvector has entries whose ratio is at most $1/E$, and similarly for the left eigenvector. Thus the ratio between two probabilities is at most $1/E^2$, which limits the spread between the best and worst pages.

Therefore, a high value of E is interesting in some circumstances when we want to distinguish only the good pages between them and leave all the bad pages to virtually the same value.

A more quantitative way to compare the different rankings is Kendall's rank correlation coefficient. Given two rankings of N objects, Kendall's coefficient is

$$\frac{(\text{number of pairs of same order}) - (\text{number of pairs of opposite order})}{\text{total number of pairs}}. \tag{9}$$

The total number of pairs is, of course, $\frac{1}{2}N(N-1)$. A value close to 0 indicates independent rankings, while a value close to 1 indicates strongly correlated rankings. Here Kendall's coefficient for PageRank compared to the free-energy rank (for both values of E) is close to 0, confirming that the free-energy rank gives information that is much different from that from PageRank. In contrast, Kendall's coefficient of the free-energy rank for $E = 0.01$ compared to $E = 3.23 \times 10^{-6}$ is 0.72, which shows that the ranking is not very sensitive to the value of E.

VI. EFFECT OF LINK FARMS

Intuitively, a node has a high entropy rank or free-energy rank if it belongs to many paths. Thus there are two ways to obtain a high entropy rank or free-energy rank: to be pointed at by good pages or to point to good pages. This is reminiscent of the HITS method [2], which computes a hub score and an authority score for every node from the dominant eigenvectors of AA^T and A^TA. The exact relationship between HITS and the entropy method remains to be investigated. Let us now see how easy it is for a malicious webmaster to artificially boost its ranking by creating a link farm, i.e., a large group of dummy pages whose structure is designed to improve the ranking of a specific page.

A typical way to increase the PageRank score of a page consists in changing the page into a good authority by adding a large number of pages, all pointing to each other and pointing to the page to be artificially increased. This technique has an interesting impact on the free-energy rank score, as shown by the following simulation.

For the piece of the Stanford web used in Sec. V we choose the page that was classified at the 200 000th rank according to the free-energy rank, for $E = 3.23 \times 10^{-6}$, and classified at the 154 325th rank according to PageRank for $\alpha = 0.9$. We then added a link farm of 100 nodes pointing to each other and to this page. This page then reached the 627th rank according

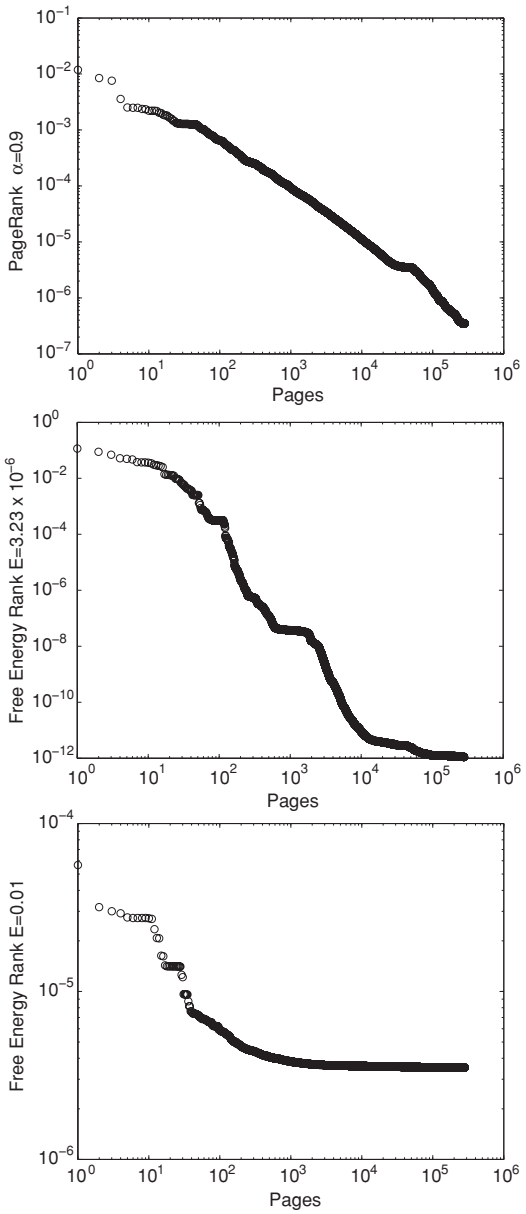

FIG. 2. PageRank and free-energy rank distributions in logarithmic scales. Top: The PageRank seems to be distributed according to a power law of slope close to -1. Middle: The distribution curve of the free-energy rank is steeper, which indicates a larger discriminating power between the best and worst pages. Here $E = 3.23 \times 10^{-6}$, which corresponds to $\alpha = 0.9$ for the PageRank. The distribution is also less regular than that at the top. Bottom: A larger value of $E = 0.01$ limits the spread of the distribution and creates an almost uniform distribution for the worst pages.

to the free-energy ranking and the 29 173th rank according to PageRank. Interestingly, the 100 new pages get an even (slightly) higher free-energy ranking than the page they are conspiring to push forward, though they get a much lower ranking than this same page for the PageRank.

Although the rank benefit is larger for the free-energy rank method, the cheating is thus easier to detect: A new plateau has appeared in the distribution of centrality around ranks 30–130 (see Fig. 3). Since the nodes in the link farm do not get as high

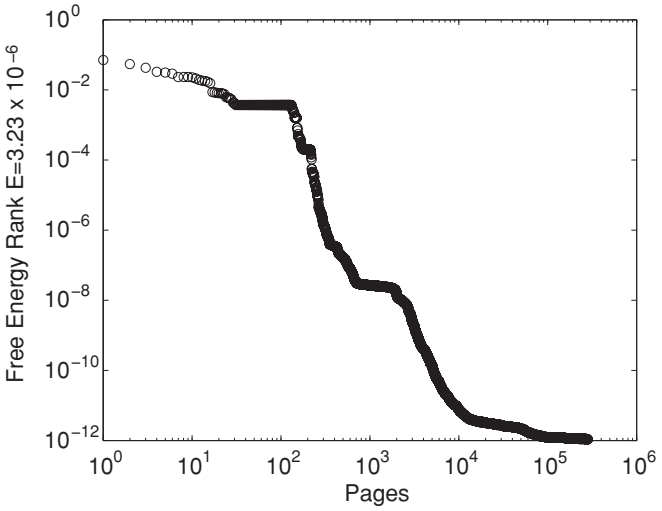

FIG. 3. Sensitivity of the free-energy rank distribution of Fig. 2 ($E = 3.23 \times 10^{-6}$) to a link farm of 100 nodes pointing to each other and to the page initially classified at the 200 000th rank. In comparison with Fig. 2 (middle), note how the 100 nodes of the link farm receive a very high ranking, thus forming a new plateau in the free-energy rank distribution.

of a score from PageRank as the page they push forward, this plateau exists, but is much less visible in the distribution curve of PageRank (around ranks 54 180–54 280).

In the HITS method we know that pointing to good pages on a topic increases the hub score of a page. Again we verify the impact of this falsification on the free-energy rank. We choose again the same 200 000th page according to the free-energy rank with $E = 3.23 \times 10^{-6}$ and make it point to the 100th best page. As a result, the 200 000th page is now 629th.

Note than in the HITS method, a page largely improves its authority score when pointed at by a link farm of 100 pages and largely improves its hub score when pointing to the best pages, but this improvement is still much less impressive than that for the corresponding falsification of the free-energy rank.

VII. EXTENSIONS

Note that the experiences above refer to the application of the centrality measures for the full graph of the web. Those centrality measures can be applied in a number of contexts. For example, in the HITS method it is usually considered that scores are computed only on the subgraph composed of those pages that contain a certain keyword and their neighbors. In this example and in other kinds of networks, such as the interbank network [7] where nodes are banks and edges are loans between them, the techniques of cheating of course do not make sense, at least not in the same way. More generally, the meaning of the different centrality measures varies according to the meaning of the network and, according to the example, one may consider one or another centrality measure to be more or less appropriate in such and such network.

We have applied the free-energy rank by endowing every nonedge ij with a certain energy U_0. Of course we could choose a nonuniform energy, which would depend on i, j, or both. This is similar to Google's replacing in PageRank the jump uniformly to any other node by a jump to any other node,

with a probability that depends on this node, in order to favor some nodes.

So far we have consider unweighted networks. On a weighted network, we can interpret the weights as energies (or even exponential if those weights are non-negative) and define a centrality measure from the stationary distribution maximizing the free-energy rate. If the graph is not strongly connected and aperiodic, then we can use the same trick again of transforming every nonedge into an edge with a certain energy U_0.

VIII. CONCLUSION

We have shown how to use the Ruelle-Bowens free-energy rate maximizing random walk on any weighted graph instead of the simple random walk in order to extract information from this graph. We applied it to centrality measures with the introduction of the entropy rank and the free-energy rank and compared it with PageRank and HITS. We compared the robustness of those centrality measures with respect to the introduction of a clique, called a link farm when it comes to the fraudulent manipulation of the web page rankings. We observed that the free-energy rank is much more sensitive to such perturbations than the PageRank and HITS. This suggests that global distribution of the free-energy rank is more sensitive to the medium-scale details of the graph and may explain why it does not appear as powerlike as PageRank.

We do not claim that Ruelle-Bowens random walk provides a better basis for centrality, only that it provides a spectral centrality that is completely different from those considered so far, with very different properties, which may be more or less suitable according to the context. We insist that this is one possible application of the Ruelle-Bowens random walk to complex networks. Every method that performs a random walk on the graph in order to analyze it, such as Markov clustering [21], a walk trap [22], stability [16,17], commute-time distance [23], and kWalks [24], could, in principle, be adapted to the Ruelle-Bowens walk. Again, the resulting algorithms would perhaps be more relevant in some cases and less so in others. The exploration of such algorithms and for which applications they are suitable opens a vast field for future research.

ACKNOWLEDGMENTS

The work of A.-S.L. was supported by the Fonds de la Recherche Scientifique. The work of J.-C.D. was supported by the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office and by the Concerted Research Action "Large Graphs and Networks" of the French Community of Belgium. This paper has benefited from discussions with the team GYROWEB at the Institut National de Recherche en Informatique et en Automatique, Rocquencourt (France) and the team Large Graphs and Networks at Université Catholique de Louvain, Louvain-la-Neuve (Belgium). In particular, Diep Ho Ngoc simplified the main proof in Sec. III.

[1] S. Brin and L. Page, Comput. Networks ISDN Syst. **30**, 107 (1998).
[2] J. Kleinberg, J. Assoc. Comput. Mach. **46**, 604 (1999).
[3] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, in *PageRank, HITS and a Unified Framework for Link Analysis, Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, USA, May 1-3, 2003*, edited by D. Barbara and C. Kamath (SIAM 2003).
[4] M. Akian, S. Gaubert, and L. Ninove, in *The T-PageRank: A Model of Self-Validating Effects of Web Surfing*, edited by C. Commault and N. Marchand, *Positive Systems, Proceedings of the Second Multidisciplinary Internation Symposium on Positive Systems: Theory and Applications (POSTS 06), Grenoble, France, August 30-September 1, 2006* (Springer LNCIS 341, 2006), pp. 1924.
[5] P. Bonacich, Am. J. Soc. **92**, 1170 (1987).
[6] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).
[7] J. T. E. Chapman, M. L. Bech, and R. J. Garratt, J. Monetary Economics **57**, 352 (2009).
[8] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White, Science **325**, 422 (2009).
[9] W. Parry, Trans. Am. Math. Soc. **112**, 55 (1964).
[10] D. Ruelle, *Thermodynamic Formalism* (Addison-Wesley, Reading, MA, 1978).
[11] L. Demetrius and T. Manke, Physica A **346**, 682 (2005).
[12] J.-C. Delvenne, e-print arXiv:0710.3972v1.
[13] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, Acta Phys. Pol. B **41**, 949 (2010).
[14] J. Gomez-Gardenes and V. Latora, Phys. Rev. E **78**, 065102(R) (2008).
[15] R. Sinatra, J. Gomez-Gardenes, R. Lambiotte, V. Nicosia, and V. Latora, e-print arXiv:1007.4936.
[16] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, Proc. Natl. Acad. Sci. USA **107**, 12755 (2010).
[17] R. Lambiotte, J.-C. Delvenne, and M. Barahona, e-print arXiv:0812.1770.
[18] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. USA **105**, 1118 (2008).
[19] J. Brown, *Ergodic Theory and Topological Dynamics* (Academic, New York, 1976).
[20] Sep Kamvar, [http://kamvar.org].
[21] S. van Dongen, Centrum Wiskunde & Informatica, Report No. INS-R001, 2000 (unpublished).
[22] M. Latapy and P. Pons, J. Graph Algorithms Appl. **10**, 191 (2006).
[23] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, IEEE Trans. Knowledge Data Eng. **19**, 355 (2006).
[24] P. Dupont, J. Callut, G. Dooms, J.-N. Monette, and Y. Deville, Département d'Ingénierie Informatique, Université Catholique de Louvain Report No. RR 2006-07, 2006 (unpublished).